Mining nearness relations from an n-grams Web corpus in geographical space

Derungs, Curdin ; Purves, Ross S

Abstract: Interacting with spatial data effectively requires systems that not only process references to locations, but understand spatial natural language. Empirical research has demonstrated that near is vague, asymmetric and context dependent. We explore near in language using Microsoft Web n-grams for expressions of the form A near*, where A are placenames referring to different spatial granularities, ranging from points of interest to large U.S. cities and * are autocomplete suggestions for placenames. Analyzing the extracted expressions requires consideration of semantic and referent ambiguity. With more than 200,000 expressions we show not only what is considered to be near at different scales, but also produce intuitive maps of nearness for different locations.

DOI: https://doi.org/10.1080/13875868.2016.1246553

Taylor & Francis
Taylor & Francis Group

Spatial Cognition & Computation
An Interdisciplinary Journal

ISSN: 1387-5868 (Print) 1542-7633 (Online) Journal homepage: http://www.tandfonline.com/loi/hscc20

Mining nearness relations from an n-grams Web corpus in geographical space

Curdin Derungs & Ross S. Purves

To cite this article: Curdin Derungs & Ross S. Purves (2016) Mining nearness relations from an n-grams Web corpus in geographical space, Spatial Cognition & Computation, 16:4, 301-322, DOI: 10.1080/13875868.2016.1246553

To link to this article: http://dx.doi.org/10.1080/13875868.2016.1246553

Accepted author version posted online: 12 Oct 2016.
Published online: 12 Oct 2016.

SPATIAL COGNITION & COMPUTATION
2016, VOL. 16, NO. 4, 301–322
http://dx.doi.org/10.1080/13875868.2016.1246553

Taylor & Francis
Taylor & Francis Group

Mining nearness relations from an n-grams Web corpus in geographical space

Curdin Derungs[a,b] and Ross S. Purves[a]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]URPP Language and Space, University of Zurich, Zurich, Switzerland

ABSTRACT

Interacting with spatial data effectively requires systems that not only process references to locations, but understand spatial natural language. Empirical research has demonstrated that near is vague, asymmetric and context dependent. We explore near in language using Microsoft Web n-grams for expressions of the form A *near* *, where A are placenames referring to different spatial granularities, ranging from points of interest to large U.S. cities and * are autocomplete suggestions for placenames. Analyzing the extracted expressions requires consideration of semantic and referent ambiguity. With more than 200,000 expressions we show not only what is considered to be near at different scales, but also produce intuitive maps of nearness for different locations.

KEYWORDS

near; spatial language; n-grams; GIR; Web

1. Introduction

Peter Fisher argued that vagueness is "in our view and understanding of everything around us, and, most profoundly, embedded in our natural language." (Fisher, 2000, p. 7). Montello, Goodchild, Gottsegen, and Fohl (2003) suggest two distinct and commonly used exemplars of vague use of spatial language: spatial relations and regions, while Fisher (2000) suggests three: geographical relations, objects and processes. These concepts broadly overlap, though Montello et al. focus on the spatial extents implied by such vagueness, yet Fisher also includes notions related to membership of categories (e.g., when does a hill become a mountain) and adds the notion of process. Nonetheless, both papers are motivated by arguing for the importance of developing geographical information systems and associated representations that capture vagueness.

The need for dealing with such vagueness assumes a particular importance in the context of systems used by non-experts. These issues are also reflected in the increasing importance given to common-sense models of geography (Egenhofer & Mark, 1995). Such calls to action are very common in geographic information science, and have motivated a wide range of research over the last

CONTACT Curdin Derungs curdin.derungs@geo.uzh.ch; ross.purves@geo.uzh.ch Department of Geography, University of Zurich, Zurich, Switzerland.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hscc.

two decades. Despite this, contemporary commercial tools dealing with geographic information do so primarily using abstractions of space limited to points, lines and polygons with sharp, well defined borders and reduce spatial relationships to buffer functions.

This is not only the case in geometric data typically used to represent administrative boundaries, but also where textual representations of space are at the forefront. Thus, for example, schema.org foresees places as being represented by either *GeoCoordinates* (effectively points) or *GeoShapes* (bounding boxes, buffered points, lines or polygons) and represents spatial relationships between places explicitly as containment. These places can range from data typically found in point information databases (such as a BusStop) through to clearly vague spatial concepts (such as a Mountain). GeoSPARQL, an extension of the popular SPARQL RDF query language represents the query predicate *spatial: nearby* as places within a metric distance of locations represented as a point, where places are represented as well-defined points or regions. Both of these currently popular models clearly neglect simple notions of spatial vagueness. In this article, we set out to explore empirically one widespread example of vagueness in geographic natural language—the use of the spatial preposition *near* to describe the relationship between two locations in text. We do so through the analysis of n-grams – contiguous sequences of tokens retrieved from a large corpus. In our case we are interested in the use of n-grams of the form *A near B* where A and B are both toponyms. N-grams are commonly used in computer linguistics in tasks including speech recognition, spelling autocorrection, query completion and machine translation (e.g., Manning, Raghavan, & Schütze, 2008; Halevy, Norvig, & Pereira, 2009). The basic underlying assumption is that a relatively simple statistical language model, derived from a suitable corpus, can provide useful information in a given context. Thus, for example, the probability of a given token occurring, given *n* previous tokens (e.g., what is the probability of *Anne* being the next token following the tokens *William Shakespeare's wife*) can be used to provide realistic query autocomplete suggestions.

Our aims in this research were threefold:

- to make a methodological contribution, demonstrating how n-gram corpora can be used to analyze spatial relationships, which goes beyond previous preliminary work (Derungs & Purves, 2014);
- to describe empirically the properties of nearness in a very large corpus, allowing us to analyze many thousands of nearness relations in simple natural language; and
- to generate models of nearness relations suitable for use in systems taking into account the vagueness in language and describe potential applications of such models.

Next we first look briefly at the increasing body of work seeking to mine textual sources for geographic information, with a particular focus on vague

spatial relations and regions, before discussing the problem of toponym recognition and resolution. An overview of our methodological approach is given, before we set out our reasons for working with n-grams and introduce the gazetteer data used. We then explain our approach to extracting and interpreting relevant n-grams, focusing on the key problem of toponym disambiguation. Finally, we present results including distances associated with near in the United States and spatial models of nearness for three U.S. cities, before exploring the extent to which useful information can be derived from our approach, and suggesting its implications for previous and future work.

2. Related work

Worboys (2001) and Fisher (2000) discussed philosophical and logical aspects of vagueness in great detail, while Montello et al. (2003) gave a comprehensive overview on cognitive issues. For a broader overview we thus refer to these three articles. Here, we narrow our focus to findings from previous empirical experiments on vague spatial relations and regions, as summarized in Table 1.

The work in Table 1 is structured using general characteristics, including the object of study, scientific domain, type of approach and short descriptions of the task or focus respectively. In the following we use these characteristics to conduct an in-depth discussion of, first, types of approaches that were applied, with a particular focus on the source of information and, second, findings on properties of vague spatial relations and regions.

2.1. Sources of information and experimental settings

Earlier work is skewed towards empirical user studies, although latterly corpus studies have become increasingly common. User studies cover a broad selection of experimental settings. In numerous studies, participants were asked to evaluate the near-, far- or closeness of given instances of spatial relations (Fisher & Orf, 1991; Robinson, 2000; Worboys, 2001; Yao & Thill, 2005). Montello et al. (2003) showed base maps of Santa Barbara to participants with the task of drawing the border (or the 50% and 100% confidence region) of Downtown. Shariff, Egenhofer, and Mark (1998) also devised an experiment containing a drawing exercise with the aim of finding differences between some 59 spatial relations between lines and polygons. A somewhat inverse approach was taken by Carlson and Covey (2005) and Stevens and Coupe (1978). In both examples spatial settings were shown, or explained, to participants, who then had to associate quantitative measures. Carlson and Covey (2005) asked for the Euclidean distance associated with spatial relations described, while in Stevens and Coupe (1978), spatial settings had to be remembered and cardinal directions recalled from memory.

User studies in cognitive linguistics and psychology show a particular interest in the relation between the spatial domain and its representation in

Table 1. Overview of papers exploring vague spatial relations.

Source	Study object	Scientific domain	Approach	Task or Focus
Fisher (2000)	spatial relation < near, close >	GIScience	user study	judging nearness
Worboys (2001)	spatial relation < near, not-near >	GIScience	user study	judging nearness
Montello et al. (2003)	spatial region < downtown >	GIScience, Psychology	user study	drawing regions
Robinson (2000)	spatial relation < near, distant, in vicinity, etc. >	GIScience	user study	judging nearness
Carlson & Covey (2005)	spatial relations < near, far, etc. >	Psychology	user study	applying distance metrics to descriptions
Shariff et al. (1998)	spatial relation <59 diff. relations between lines and polygons >	GIScience	user study	drawing described spatial relations
McDonough et al. (2003)	spatial relation < in, on >	Cognitive Linguistics	user study	classifying spatial settings
Skoumas et al. (2013)	spatial relation < near, in, south, etc. >	GIScience	corpus study	distance and direction from instances of spatial relations in text
Stevens & Coupe (1978)	spatial relation < direction >	Psychology	user study	judging directions between places and points
Piwek et al. (2008)	spatial relation < proximal, distal >	Cognitive Linguistics	user study	references to proximate and distant objects
Yao & Thill (2005)	spatial relation < near, far >	GIScience, Computer Science	user study	judging nearness
Schockaert et al. (2008)	spatial relation and region	GIScience, Computer Science	corpus study	distance of nearness relations as used in the Web
C. B. Jones et al. (2008)	spatial region	GIScience	corpus study	collocation of region and place names in Web documents
Hollenstein & Purves (2010)	spatial region	GIScience	corpus study	use of spatial regions in user generated content
Wallgrün et al. (2014)	spatial relation < near, close, next >	GIScience	corpus study	distance and travel time of nearness relations between POIs in Web documents
Derungs & Purves (2014)	spatial relation < near >	GIScience	corpus study	probabilities of near relations from Web n-grams for cities in UK

language and cognition. Piwek, Beun, and Cremers (2008) study correlations between the use of near and distant demonstratives in language and pragmatic characteristics of the reference objects, such as availability or importance. McDonough, Choi, and Mandler (2003) compare prelanguage classifications of spatial settings across different cultures with classifications given by adults.

Most corpus studies aim to identify instances of spatial relations or regions in text, tags or n-grams. Aggregations of instances are either presented in the form of descriptive statistics of measurements associated with spatial relations, e.g., distance (Derungs & Purves, 2014), distance and travel time (Wallgrün, Klippel, & Baldwin, 2014) or characteristics of reference objects (Schockaert, de Cock, & Kerre, 2008), or as spatial distributions usually in the form of density surfaces (Hollenstein & Purves, 2010; C. B. Jones, Purves, Clough, & Joho, 2008; Skoumas, Pfoser, & Kyrillidis, 2013).

2.2. Properties of vague spatial regions and relations

Variations in interpretations of natural language containing vague spatial relations or vague regions have typically been explained through a detailed analysis of the impact of context (e.g., Fisher & Orf, 1991; Montello et al., 2003; Wallgrün et al., 2014; Worboys, 2001). Carlson and Covey (2005), for instance, found that the size of reference objects determines the distance associated with nearness relations. Large objects may be further away and still considered near. Stevens and Coupe (1978) noted that people employ a storage-saving strategy by remembering primarily the structure of spatial settings and then infer details. Complex hierarchical structures have thus proved to lead to distortions in the inferred information. Yao and Thill (2005) incorporated a comprehensive list of context parameters, such as type of and familiarity with activity, user experience or demographic information, to compute a regression model that allows measurement of the influence of context on nearness judgements. In their case study, activity related context information had the greatest predictive power. Worboys (2001) suggests that nearness is an example of a similarity relation with weakened formal properties, such that, for example, symmetry is not guaranteed. Depending on the contextual setting (e.g., uneven distribution of population), *A near B* does not have the same meaning as *B near A*. Vagueness is often conceptualised and represented as fuzziness. Montello et al. (2003) found the region of Downtown Santa Barbara to have a fuzzy boundary after overlaying multiple instances of drawn borders.

Fuzziness is also represented by the density surfaces computed in Hollenstein and Purves (2010), C. B. Jones et al. (2008) and Skoumas et al. (2013). However, in these approaches fuzziness is a product of the method used, namely kernel-based approaches, rather than an inherent characteristic of the results. From this overview of empirical work on vague spatial relations and regions, it is apparent that case studies often have a particular focus on one

scale (often environmental space *sensu*; Montello, 1993) and are limited in terms of the amount of data that is compiled. Furthermore, empirical work typically focuses on judging post-hoc relationships between given objects, rather than analysing the use of vague spatial language *in situ*. Corpus studies, which offer considerable potential for such research remain underrepresented, and to date have mostly focused on describing vague spatial regions rather than spatial relationships.

2.3. Toponym grounding and ambiguity

A key step in analysing spatial relationships from text corpora is grounding toponyms – that is to say associating placenames with unique sets of geographic coordinates. Hill (2006) argued that "georeferencing by placename (aka feature name) is the most common form of referencing a geographic location [...]" (p. 91), which in turn implies that much information is not associated with explicit geographic coordinates. The link between information and location is thus often made through toponyms. However, associating toponyms found in text with a unique location first requires their recognition and, in a second step, disambiguation (Leidner & Lieberman, 2011). Toponym recognition is often accomplished through gazetteer lookup, where each word in a text is compared to entries in a list of toponyms (i.e., gazetteer). Matches to the gazetteer are tagged as candidate toponyms (Clough, 2005). Candidate toponyms are then disambiguated, a task which must deal with two important types of ambiguity. Semantic ambiguity is very common in toponyms since they are also used in other senses (i.e., there are cities in the United States with names such as Kathleen, Dennis, and Home).

Furthermore, associating toponyms with unique locations is challenging because of referent ambiguity (there are approximately 60 places called Springfield in the United States) (Leidner, 2008). Buscaldi (2011) distinguishes three approaches to toponym disambiguation, namely *map-* and *knowledge-based* and *data-driven*. Map-based approaches require text sequences containing several toponyms, yet knowledge-based approaches are dependent on additional context. The most prominent knowledge-based approach, often referred to as default disambiguation, resolves referent ambiguity by only considering a single default location (for example the most populous) with respect to a toponym (e.g., Amitay, Har'El, Sivan, & Soffer, 2004) and ignoring all other candidates. Data-driven approaches apply machine learning techniques and are thus dependent on gold standard training data (e.g., Martins, Anastácio, & Calado, 2010; Santos, Anastácio, & Martins, 2015), which in turn requires contextual information for annotators. From this short introduction to toponym disambiguation it is obvious that disambiguation of n-grams is only possible to a limited degree because the texts are extremely short and, due to lack of surrounding context, creation of a gold standard is difficult.

3. Analyzing nearness using n-grams

3.1. Overview

Our basic approach to analyzing nearness was to associate phrases of the form A near B with probabilities returned by an n-gram service, where A and B were toponyms. Because toponyms are inherently ambiguous, we also grounded individual toponyms to unique locations taking account of both referent and semantic ambiguity. To deal with referent ambiguity we associated each toponym with a single, unique location, yet for semantic ambiguity we assigned a weight to each toponym, which was then used during the aggregation or visualization process. Figure 1 shows example results illustrating our approach at two contrasting scales: locations *near Central Park* within New York, using point of interest (POI) data, and *near New York* using data for populated places over the entire United States. Having generated such grounded data, we could then aggregate and weight distances or probabilities for all found pairs to generate box plots illustrating the distribution of distances and probabilities related to a particular use of near.

Next, we describe and characterize, first, the source data we worked with in our analysis, before describing our methods in more detail.

3.2. Data

Before investigating the occurrence of phrases such as A near B, where A and B are toponyms, a number of questions can be posed. The first of these concerns the choice of phrase used to mine nearness relationships. Our approach was to find a simple phrase that was likely to be representative of the spatial

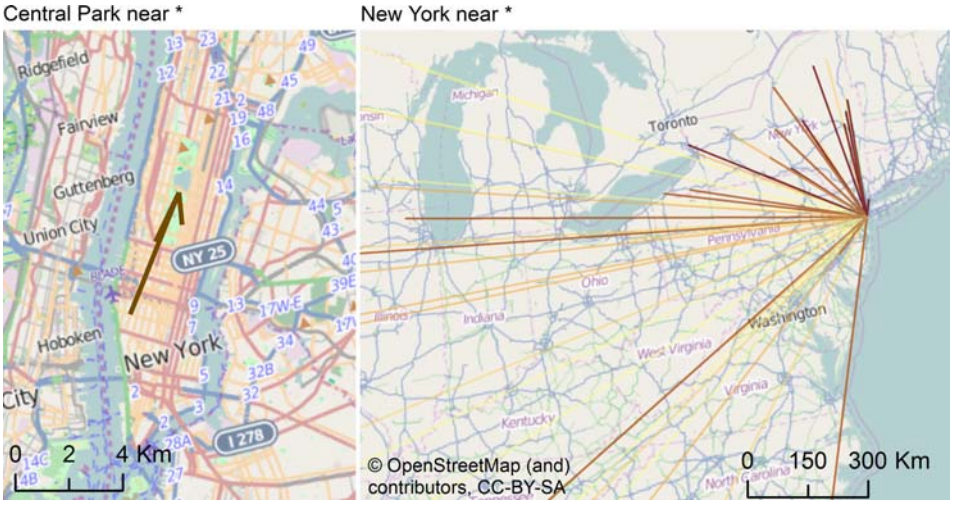

Figure 1. Near relations for Central Park near * ($n = 4$) and New York near * ($n = 50$). Bright colors in the New York example represent high degrees of semantic ambiguity.

Table 2. Overview of selected full-text corpora.

Corpus	Number of documents	Source
ClueWeb12	700 million	boston.lti.cs.cmu.edu/clueweb12/
Wikipedia	4.4 million	corpus.byu.edu/wiki/
Alpine Journal	13.2 thousand	textberg.ch/site/de/korpora/
GloWbE	1.8 million	corpus.byu.edu/glowbe/

relationship of interest, and sufficiently common in natural language (c.f. Wallgrün et al., 2014). We performed an initial study of the use and form of nearness phrases in two full text corpora using preexisting query mechanisms and part of speech tags. GloWbe (corpus.byu.edu/glowbe) is a corpus containing some1.8 million English Web pages and 1.9 billion words. Here we noted that phrases of the form *near B*, where B was a toponym were among the most common collocates of the form *near* *. Very typical collocates of the form * *near* were nouns, with the most common forms being generic nouns. Thus, phrases such as mountains near Glasgow were common in this corpus. To investigate how often the nouns preceding near took the form of toponyms, we used another corpus, the yearbooks of the Alpine Club, which consist of some 6.3 million words (www.textberg.ch/ReleaseNotes/README_Release_151_v01.htm), where full part of speech tagging and named entity annotation was available. In this corpus we found that of around 1,700 instances of near, 160 took the form *A near B* where A and B were both named entities (and thus, likely to be toponyms). Such full text corpora allow us to analyse the nature and appropriateness of phrases containing spatial relationships, but they are still relatively rare, especially when we focus on simple constructions (where dependencies are straightforward). Table 2 gives an overview of four well-known and commonly used full text corpora, including the two described previously.

Analyzing nearness relations in these corpora would be possible, but would require part of speech tagging and named entity recognition, including toponym resolution. However, having established that phrases of the form *A near B* occurred with a frequency of around 10% in a corpus where such part of speech tagging was available, we chose another approach, the use of n-grams, where much larger volumes of data are available. N-grams corpora are essentially look up tables, allowing estimations of the probabilities of particular phrases (either in exact forms or as represented by wild cards), generated from very large corpora. We used the Microsoft Web N-Gram Service[1] in this research, which is sourced from more than 100 billion Web pages (Wang, Thrasher, Viegas, Li, & Hsu, 2010), and thus more than hundred times larger than the largest corpora in Table 2. The underlying assumption is that, having established that phrases of the form A near B occur, many more instances will occur within this very large collection. Querying the service requires registration, and a user token permits queries related to body, anchor text, and

1. weblm.research.microsoft.com/info/index.html

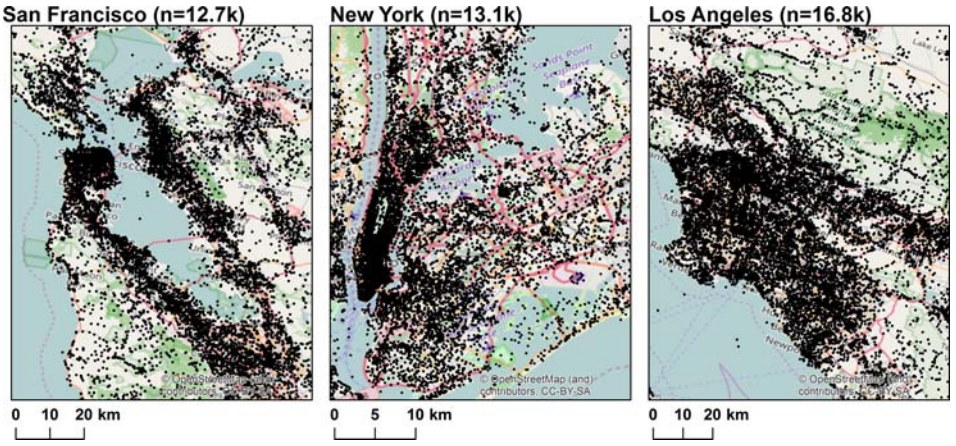

Figure 2. Maps of the geonames.org POI distribution (black dots) in three U.S. cities.

titles of Web pages indexed by Bing in 2013 in the *en-us* market to be made. In our work, we looked at body text, and used three different functionalities of the service. First, joint probabilities, represent the probability of a submitted set of tokens occurring in the n-gram corpus:

$$P(w_1, \ \ldots \ , w_m) = \prod_{i=1}^{m} P(w_i | w_1, \ \ldots \ , w_{i-1})$$

where w_i is a given token. Such joint probabilities can effectively be considered equivalent to a frequency. Because low-frequency n-grams may not have been observed in a given corpus, computational linguistic methods such as smoothing are used to approximate joint probability values (e.g., Jurafsky & Martin, 2000). Second, conditional probabilities, expressed as $P(w/c)$, where w is a word and c a sequence of words preceding it, allow us to explore how likely a particular word is given some preceding context (e.g., the probability of sun being the next token following the tokens severe weather warning). Third, it is also possible to query the service with a given sequence s and retrieve the set of words $w_1, \ \ldots \ , w_n$ most likely to follow this sequence, ranked as a function of conditional probability. To both generate queries and localise toponyms associated with near, we used gazetteers. Because we effectively worked at two scales: both within individual cities and across the United States as a whole, we used two different gazetteers. Within cities we used points of interest (POI) from GeoNames for three U.S. cities, namely New York, San Francisco, and Los Angeles (Figure 2). We considered all named entities located within the bounding box of each of the three cities as POIs. Across the United States, we used a hierarchical gazetteer containing some 30,000 populated places and their populations, made available by the U.S. census bureau[2] in 2010 (the last year with population counts). It is

2. www.census.gov/geo/maps-data/data/gazetteer2010.html

worth noting that all of the data sources we used here are publicly available, and it should thus be more straightforward to replicate our methods and reproduce our results. However, it is important to note that the Microsoft n-grams service is effectively a black box, and that results may change if; for example, the approach to smoothing is varied in the future.

4. Methodology

In a preliminary study (Derungs & Purves, 2014) we generated a matrix of joint probabilities for n-grams of the form A near B for all toponyms contained in a given UK gazetteer. However, this work demonstrated that such joint probabilities were highly prone to ambiguity, and that filtering was necessary to remove many spurious probabilities. Therefore took a different approach, and developed a recursive algorithm to extract likely combinations of place names from the n-grams service, before further filtering these for ambiguity where additional contextual information was available.

Our algorithm can be described as follows. Firstly, for every toponym *A* in the gazetteer we form a sequence *s* of the form *A near*, which is submitted to the n-grams service. Note that many toponyms consist of more than one token, and the maximum length of sequence that can be submitted to the service is four, allowing us to potentially query for all toponyms up to and including tri-grams (e.g., New York City). We then retrieve up to 100 suggestions for the tokens most likely to complete this sequence $w_1, \ldots, w_n$. These tokens are then iterated through to check for gazetteer matches. Here, three cases must be distinguished:

1. A token is not contained in the gazetteer and is discarded.
2. A token is an exact match with a toponym in the gazetteer and is associated with an ID.
3. A token is a partial match with a toponym in the gazetteer (e.g., *New* is a partial match with *New York*, *New Hampshire*, etc.) In this case, the algorithm is then called recursively, using as a sequence initial sequence *s* with the candidate token concatenated (e.g., *A near* w_1).

It is important to note that a token may be both an exact and a partial match and thus associated with multiple tokens. The process is continued until all candidate sequences containing four tokens or less have been queried and a set of IDs retrieved for candidate toponyms B forming sequences of the form A near B (note that given our limitation on sequence length of four, then the limiting lengths of toponyms are that either A and B may both consist of two tokens, or A or B may consist of three tokens and one token, respectively, giving a total length of five).

If all toponyms were unique and not associated with any other meanings, then our approach would essentially allow us to identify pairs of unique referents to locations, which could be directly associated with coordinates. For

our POI data, where no further contextual information is available, and where toponym granularities are typically fine, we assume this to be the case, and not to have a significant influence on our results.

However, where additional contextual information is available, we undertake both referent and semantic disambiguation. For referent disambiguation we use a default sense approach, using population as contextual information, which typically has high precision where population is unequally distributed between multiple references to a single toponym. Here, only the location associated with the most populated instance of toponym is retained. Dealing with potential semantic ambiguity is more complex.

To do so, we first calculated the joint probabilities for all toponyms in isolation (bearing in mind that toponyms may be made up of sequences of up to three tokens) and plotted these probabilities against the highest population associated with the toponym (because through our referent disambiguation process only these locations were considered further). Our basic hypothesis was that more populated places will also appear more often in Web documents, and thus have higher joint probabilities. We hypothesized that toponyms with high joint probabilities but low populations are often semantically ambiguous, and we assigned individual toponyms with a weight according to their distance from a linear regression calculated for population rank against joint probability rank as follows:

$$\epsilon_i = pop_i - (\alpha + \beta P_i)$$

Where ϵ_i is the residual value associated with toponym i; pop_i is the population rank associated with toponym i; P_i is the joint probability rank associated with toponym i; and α and β are the intersect and gradient, respectively, associated with the fitted linear regression.

Having calculated the residual we assigned a *semantic ambiguity weight* to each toponym instance. Where the residual was positive (that is to say where the population was more than would have been expected by looking at the joint probability), we assign a weight of one. Where the residual is negative (that is to say, given the population value, the toponym appears to be used more than expected), we assign a weight between 0 and 1, with the weight decreasing as the residual value deviates more from the regression line.

For both POI data and populated places we then calculated a range of measures to illustrate the use of near in our corpus. First, we generated summary statistics of the distances associated with near relations for grounded toponyms. To allow comparison with simple baselines, we also calculated distributions for random sets of toponyms and nearest neighbors. In the case of nearest neighbors (where we identified for the phrase *A near* the 50 nearest toponyms), we generated both 1 and 10 nearest neighbors randomly selected from the candidate set of 50. Where populated places are being explored, our default referent disambiguation selects candidate toponym locations for

analysis, yet the semantic ambiguity weight is used to calculate weighted distance and probability distributions visualized as box-whisker plots (the semantic ambiguity weight of an A near B relation is simply the product of the weights of A and B).

In a second analysis, we concentrated on U.S. populated places only and tested A near B relations for asymmetry, assuming joint probabilities are higher and distances smaller for relations where population(A) $<$ population(B).

In a final analysis, we visualized nearness for three example U.S. cities by calculating kernel line densities based on the complete set of nearness relationships found for all POIs in a city. For comparison, near maps are shown in combination with random and nearest neighbor maps (i.e., line density contour maps for random and nearest neighbor relations between POIs).

5. Results and interpretation

In Next we explore some general properties of toponyms in our corpus, and comment on their implications for ambiguity. Then, we demonstrate typical distances associated with nearness, and discuss asymmetric use of nearness as a function of population. Finally, we generate near maps for three U.S. cities using n-grams, and compare these to representations based on nearest neighbors and a random baseline.

5.1. Place name characteristics in n-grams

Figure 3 shows the distribution of joint probability values for populated places in the United States and POIs in our three U.S. cities.

All four distributions show similar patterns following Zipfian distributions (Zipf, 1935), with a small number of candidate toponyms having much higher probabilities. The labels attached to these candidate toponyms show clearly that most of these frequent words are highly semantically ambiguous. Home, time, may, night and good are frequent English words, and their prevalence in these distributions suggests (unsurprisingly) that joint probability values for candidate toponyms in isolation from the n-gram corpus are strongly

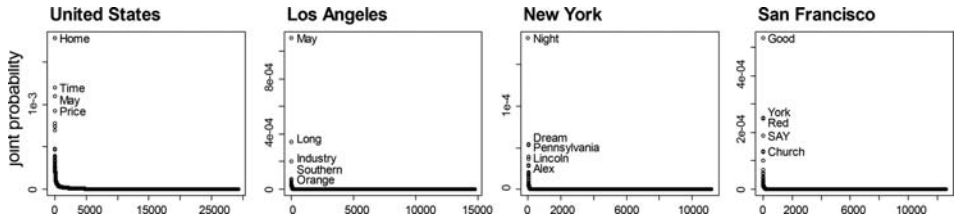

Figure 3. Distribution of n-gram joint probabilities (i.e., frequency of occurrence) of place names in the United States (Source: census bureau) and POIs in U.S. cities (Source: geonames.org).

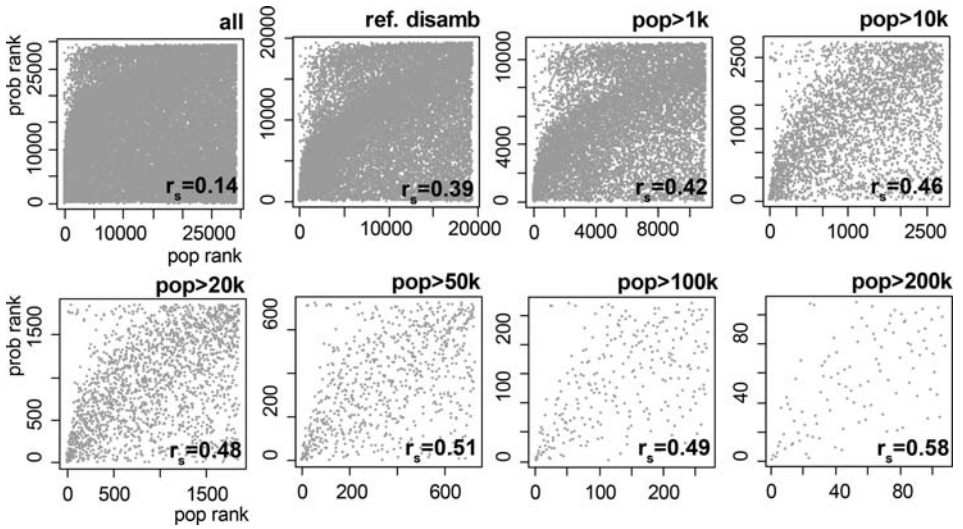

Figure 4. Correlation between population rank and joint probability rank of U.S. place names before filtering, and after default referent disambiguation and filtering according to overall population.

influenced by ambiguity. To test this hypothesis, we correlated joint probability values for U.S. populated places with population counts (Figure 4).

Three observations can be made on the basis of Figure 4. First, correlation values for population rank and joint probability rank are low, where ambiguity is ignored (0.14). Second, applying a simple baseline default referent disambiguation based on population increases the correlation substantially to 0.39. Third, by filtering out less populated places the correlation further increases to a maximum value (for the 100 populated places with population of more than 200,000) of 0.58.

Summarizing these results, it is possible to make three more general points:

- Raw joint probabilities of candidate toponyms have typical Zipfian distributions, and are dominated by semantically ambiguous toponyms.
- By filtering candidate toponyms for referent ambiguity and more populated places, we obtain strong correlations between population rank and joint probability rank, suggesting that filtered toponyms are likely to refer to locations (because we assume that Web page coverage is somewhat correlated with population).
- Because no semantic disambiguation has been carried out, it seems likely that more populated places are, overall, less likely to be semantically ambiguous with very commonly occurring words.

5.2. How far is near?

Figure 5 summarizes distances for the three POI gazetteers for New York, Los Angeles and San Francisco, where no additional disambiguation was carried

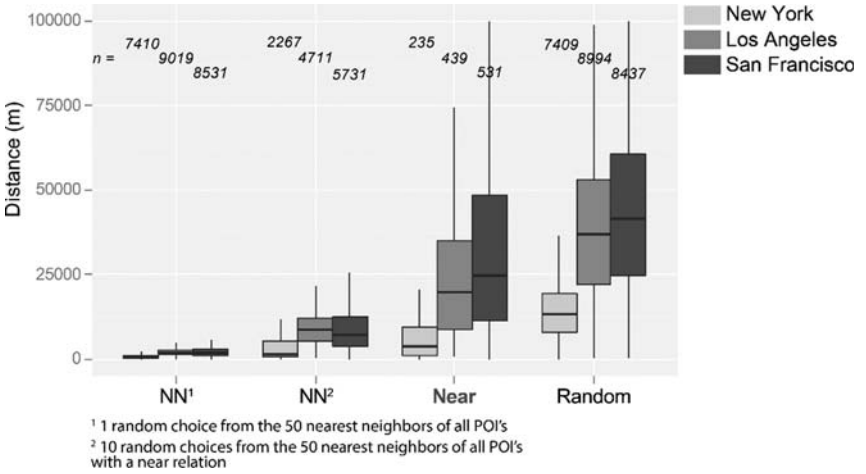

Figure 5. Comparison of nearest neighbor (NN), near and random distances between POIs in three U.S. cities.

out. The difference in median distance between the three cities, which is independent of measure (nearest neighbor, near or random) simply reflects the size of the study area with POIs in New York in particular distributed over a smaller spatial extent (c.f. Figure 2). The pattern evolving from comparing the four measurements, on the other hand, is similar for all three cities. Near distances are greater than distances between nearest neighbors but less than random distances. This result suggests that, even without explicit referent or semantic disambiguation our POIs are particular enough within relatively small bounding boxes to generate bulk results that reflect the use of near in a particular context. Thus, the median near value in New York City of 4 km in comparison with the 19 km and 25 km found for Los Angeles and San Francisco, respectively, may suggest some properties of the urban space itself, first with respect to its extent (i.e., near is likely to be a shorter distance in a smaller bounding box) and second its properties (i.e., the walkability of New York may also be reflected in shorter characteristic near distances).

However, comparing nearness across cities is difficult, since study area clearly has a strong initial influence on distributions. By comparing relative distances, with respect to mean random distance for a given study area, it is possible to gain more insight into the overall behavior of nearness (Figure 6) in this respect.

First, things in New York are not just nearer because of the reduced spatial extent and the resulting density of POIs—it appears that near really does mean something quite different in a New York context to San Francisco or Los Angeles. Once again, we suggest this might have something to do with the very different forms of these cities. Second, this suggests that analyzing nearness cannot be reduced to a function of toponym type (in this case POIs), but

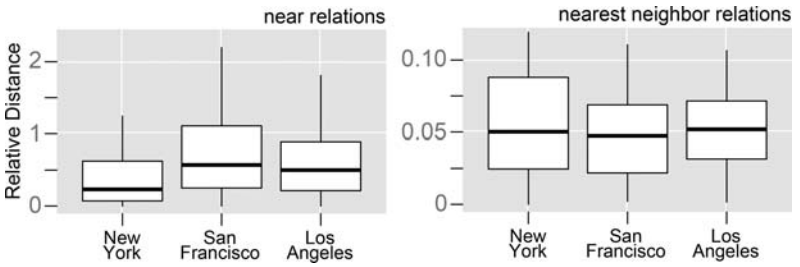

Figure 6. Relative near-distances for POIs in U.S. cities and populated places in the United States. Distances are represented as the ratio of mean random distances, i.e. a relative distance of 1 is equal to mean random distance.

must be investigated in more depth in the individual setting of a city being investigated.

Figure 6 shows similar results to Figure 5 for all populated places in the United States. Here, the first point of note is the similarity between distributions for raw near values and a random distribution. This suggests that without disambiguation, unlike for POI data, no meaningful information on the use of near at this scale can be extracted. By applying referent disambiguation and filtering for populated places with more than 10,000 inhabitants, near distances with a median of 250 km are found, however this approach to filtering reduces the dataset to some toponym 3,000 pairs. By using referent disambiguation and our semantic ambiguity weight, it is possible to retain a much larger sample ($n = 14{,}000$) and we retrieve very similar median near distances. Once again, after disambiguation, the overall pattern is unsurprising but informative characteristic distances associated with nearness are less than those from a random distribution, but greater than nearest neighbors. These results demonstrate that by using n-grams we can derive characteristic near values for very large numbers of locations (14,000 here).

One final remark should be made with respect to Figure 7. The impact of filtering on the distance statistics clearly demonstrates the impact of referent and semantic toponym ambiguity at these scales. However, at city scales, that is

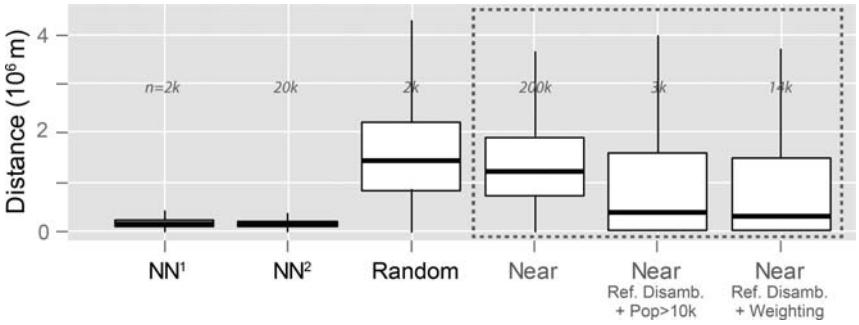

Figure 7. Comparison of nearest neighbour, different near and random distances between place names in the U.S. (Superscript numbers are described in Figure 5).

Table 3. Count of A near B relations, separated into relations where A has smaller population than B and vice versa after referent disambiguation.

pop A < B	8900
pop A > B	4700

in our POI data, where no contextual information suitable for disambiguation was available, we observed that expected orderings of distances were obtained without explicit disambiguation. This suggests that, in this setting, the use of context in the form of querying for completions of a sequence A near is already sufficient to handle some ambiguity, yet at a country level this is insufficient.

5.3. Asymmetry of near

By analyzing resolved near relations for populated places in the United States, it becomes clear that the suggested populated places B, completing a sequence A near are likely more populous than A. This is illustrated in Table 3, which shows counts of A near B relations after referent disambiguation for both possible cases. Nearly twice as many instances of relationships of A near B where the population of A is less than B are returned. For 1-gram (single token) toponyms, it is also possible to calculate the conditional probabilities for these two cases (Figure 8). Once again, we find clear evidence of asymmetry. Thus, A near B relations are not only more numerous for cases where A is less populous than B, they are also more probable.

5.4. Near maps

After focusing on expected properties of near relations, such as the relation to distance, the impact of context, asymmetry and, importantly, the role of

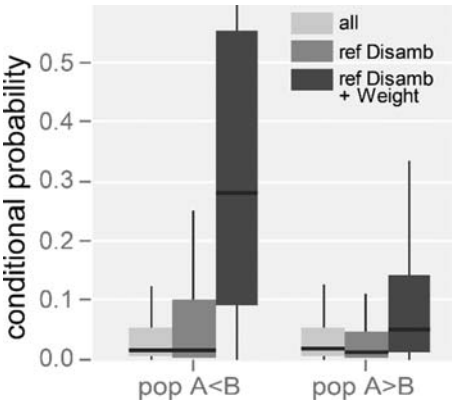

Figure 8. Conditional probability of A near B relations, where on the left, A has smaller population than B, on the right, vice versa.

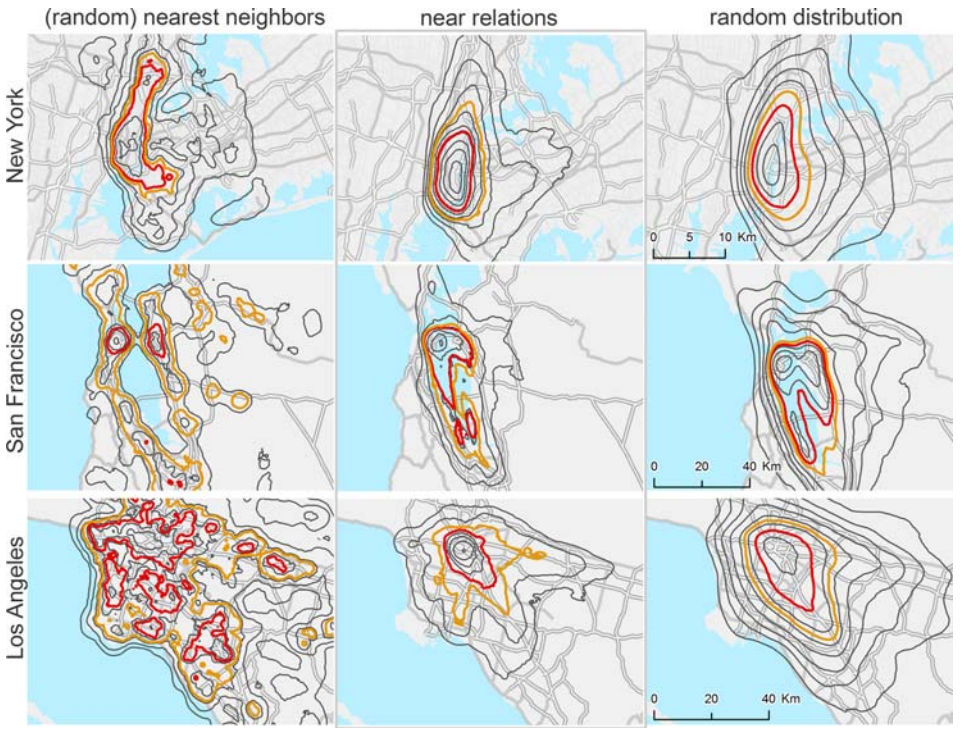

Figure 9. Contour maps for line density surfaces calculated for nearest neighbors, near and random relations between POIs in three cities. The orange and red contour line highlight the top 10% and 5% of density values, respectively.

toponym ambiguity, we used n-grams to map nearness. In Figure 9 we use all near relations identified between POIs in New York, Los Angeles, and San Francisco to compute line density surfaces where higher densities identify locations that are frequently related through near relations. To allow comparison we also visualize densities for nearest neighbors and random relations. Densities are visualized using contour maps.

Similar to the aggregate results shown in Figure 5, the contour maps first allow us to observe that nearness occurs at distances intermediate to those of nearest neighbors and random distances. However, the maps also facilitate exploration of the spatial patterns of these relationships. Nearest neighbor densities are patchy and often fail to bridge spatial entities that are supposedly interlinked, such as Times Square and the East Side of Manhattan, San Francisco, and Oakland, or the different parts of Downtown Los Angeles.

Random distributions, on the other hand, smooth out many of the finer spatial structures. New York and Los Angeles are represented as merely concentric surfaces, though in San Francisco the gap in relations, caused by the San Francisco Bay is still visible in the top 5% of densities (orange contour). Near maps, in contrast, are intermediate, allowing us to form initial hypotheses about the use of nearness spatially. For example, the use of near in New York is

concentrated and aligned with Manhattan Island, yet in Los Angeles there is clear tendency to align with the axes of the major highways centered around Downtown, suggesting the importance of road transport in the definition of nearness in this setting. Furthermore, they are straightforward to generate, along with other potential cases (e.g., our baselines here) and well suited to comparative work.

6. Concluding discussion

In the introduction we argued that contemporary commercial GIS still merely represent geographic information as points, lines or polygons and thus fall short in accounting for vagueness (e.g., Williamson, 1994). Additionally, we identified a research gap from summarizing empirical work on spatial vagueness in cognitive linguistics, psychology and GIScience; corpus studies are underrepresented and mainly focus on vague spatial regions. Corpus studies, however, have the potential to cover large spatial extents, different scales, incorporate various types of contextual information and to gain large sample sizes. In this study we cover this gap by using a Web n-gram corpus (Wang et al., 2010), i.e., a word sequence index summarizing some hundred billion webpages, to explore the meaning of near at a geographic scale (e.g., Montello, 1993).

Our first aim was to show how n-gram corpora might be used to explore spatial relationships. Our approach focused on the use of a simple phrase, backed up by an initial study in richer (but smaller) full text corpora. Simple disambiguation approaches allowed us to extract large numbers of nearness relations at multiple scales, and extending our method to deal with other spatial relationships (with the proviso that n-grams currently have a maximum length of five) would be straightforward. We also wished to explore the empirical properties of nearness in a very large corpus. Our corpus-based approach allowed us to easily investigate nearness across geographic scales, namely in individual cities and the entire United States. At city scales, returning the most probable tokens to complete a sequence appears to already perform adequate disambiguation, yet on a U.S. scale, it was necessary to apply existing methods for referent disambiguation, and develop new approaches to dealing with semantic ambiguity. Wallgrün et al. (2014) investigated nearness between POIs in cities using mined relationships from blogs, but for a much smaller dataset.

Our results accord well with previous work, showing that the metric distance associated with near is prone to variation, but differs from simply choosing the nearest neighbor. This result was robust across scales, after we had performed disambiguation at the U.S. scale (Figure 5 and Figure 7). The incorporation of three cities allowed us to test if the use of near differs in different geographic settings. We found that near in New York is almost exclusively restricted to Manhattan, leading to shorter, or walkable near distances, whereas near in Los

Angeles and San Francisco appears to relate to driving distance (Figures 5, 6 and 7). Interestingly, we also found clear evidence for the oft-stated asymmetry of nearness (c.f. Schockaert, De Cock, & Kerre, 2011; Worboys, 2001), with A near B being more probable if A is less populous than B (Figure 8). Finally, our nearness maps (Figure 9) allow us to go beyond descriptive statistics and explore the spatial structure of nearness relations in individual cities, leading to new hypotheses for the interplay between geographic setting and the use of near. Examples are the impact of road networks in Los Angeles, the bay in San Francisco or the relative isolation of Manhattan.

Our final aim was to generate nearness models suitable for use in systems better able to deal with vague concepts such as near. Based on our results, we can make several proposals for future work to evaluate the efficacy of such approaches. First, and straightforwardly, we can generate varying thresholds for nearness at the national scale for many thousands of locations, which could be used to vary query distances in search (c.f. R. Jones, Zhang, Rey, Jhala, & Stipp (2008)). Secondly, for three exemplar cities we produced spatial models of nearness, suitable for use in approaches seeking to reason using a fuzzy representation of nearness (c.f. Schockaert, De Cock, & Kerre, 2009). Other possible exploitations of our approach include selection of more likely referents from candidate toponyms pairs in producing realistic natural language phrases, using a population-based model (c.f. Hall, Jones, & Smart, 2015). It is, however, important to note that all of these proposals would require implementation and, more importantly, evaluation, to demonstrate their efficacy. However, using n-grams allows us (in contrast to previous empirical work) to generate sufficient data volumes such that typical evaluation approaches used in information science might now be possible (Pustejovsky et al., 2015). Nonetheless, our work, and more generally research exploring geographic problems through n-grams is subject to numerous limitations. The first, and most important, of these is the impact of ambiguity on toponym grounding, and thus the extraction of meaningful relations from the data. In other corpus-based studies, disambiguation appears to have been neglected Skoumas et al. (2013), carried out manually Wallgrün et al. (2014) or to have used contextual information not available for n-grams C. B. Jones et al. (2008). A key side effect of toponym disambiguation is its impact on sample size—from an initial 200,000 instances of near relations, we filtered some 90% before we could detect patterns that make sense. We caution others planning research using n-grams to carefully consider the implications of ambiguity—ignoring it is clearly not an option, without clear evidence to the contrary. Initially we planned to use joint probability values as a key measure.

However, these may have undesirable, or at least unknown, properties if below a certain threshold value. Joint probabilities of unseen n-grams are for instance calculated through smoothing (Jurafsky & Martin, 2000).

This is a serious limitation for geographic analysis, since combinations of toponyms and spatial relations, as used in this study, are not among the most frequent word sequences in the Web. Other n-gram studies usually focus on more prominent notions, such as grammar use or cultural trends (e.g., Michel et al., 2011). We thus primarily used lists of tokens completing sequences to generate samples, before calculating probabilities. This guarantees that we only explore more commonly used word sequences in the underlying corpus, but again limits sample sizes, especially compared to studies where whole n-spatial-features x n-spatial-features matrices are tested (e.g., Derungs & Purves, 2014; Worboys, 2001).

Finally, it is important to remember that n-gram corpora have maximum length of token sequence. In our case the limitation of five, although sufficient for spelling correction or machine translation (e.g., Brown, Desouza, Mercer, Pietra, & Lai, 1992), is very short for spatial relationships, primarily because of multi-token toponyms. Furthermore, it excludes the possibility of retrieving additional useful contextual information, such as that related to an activity (c.f. Yao and Thill, 2005). Despite all these limitations, we believe that n-grams provide a potentially novel and very rich source of data for the use of spatial relations in natural language, at least as practiced on the Web, with coverages and at scales that were previously out of reach, and complimenting ongoing research using full-text corpora.

References

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging web content. In *Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval* (pp. 273–280). New York, NY: ACM. Retrieved from http://doi.acm.org/10.1145/1008992.1009040

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, *3*(2), 16–19.

Carlson, L. A., & Covey, E. S. (2005). How far is near? inferring distance from spatial descriptions. *Language and Cognitive Processes*, *20*(5), 617–631.

Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the internet. In C. B. Jones & R. S. Purves (Eds.), *Proceedings of the 2005 Workshop on Geographic Information Retrieval* (pp. 25–30). New York, NY: ACM. Retrieved from http://doi.acm.org/10.1145/1096985.1096992

Derungs, C., & Purves, R. S. (2014). Where's near? Using web n-grams to explore spatial relations. In K. Steward, E. Pebesma, G. Navratil, P. Fogliaroni, & M. Duckham (Eds.), *Giscience 2014: 8th International Conference on Geographic Information Science* (pp. 23–26). Department of Geodesy and Geoinformation Vienna University of Technology. Retrieved from http://www.giscience.org/download/proceedings/GIScience2014EA.pdf

Egenhofer, M. J., & Mark, D. M. (1995). Naive geography. In A. U. Frank & W. Kuhn (Eds.), *Spatial information theory a theoretical basis for GIS: International conference COSIT '95 Semmering, Austria, September 21–23, 1995 Proceedings* (pp. 1–15). Berlin, Heidelberg: Springer. Retrieved from http://dx.doi.org/10.1007/3-540-60392-1_1

Fisher, P. F. (2000). Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, *113*(1), 7–18.

Fisher, P. F., & Orf, T. M. (1991). An investigation of the meaning of near and close on a university campus. *Computers, Environment and Urban Systems*, *15*(1), 23–35.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, *24*(2), 8–12.

Hall, M. M., Jones, C. B., & Smart, P. (2015). Spatial natural language generation for location description in photo captions. In I. S. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, & S. Bell (Eds.), *Spatial Information theory: 12th International conference, COSIT 2015, Santa Fe, Nm, USA, October 12–16, 2015, Proceedings* (pp. 196–223). Cham, Switzerland: Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-23374-1_10

Hill, L. L. (2006). *Georeferencing: The geographic associations of information*. Cambridge, MA: MIT Press.

Hollenstein, L., & Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, *1*(1), 21–48.

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, *22*(10), 1045–1065.

Jones, R., Zhang, W. V., Rey, B., Jhala, P., & Stipp, E. (2008). Geographic intention and modification in web search. *International Journal of Geographical Information Science*, *22*(3), 229–246.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Tech. Rep.). Upper Saddle River, NJ: Prentice Hall.

Leidner, J. L. (2008). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Boca Raton, FL: Dissertation.com.

Leidner, J. L., & Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, *3*(2), 5–11.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge, UK: Cambridge University Press.

Martins, B., Anastácio, I., & Calado, P. (2010). A machine learning approach for resolving place references in text. In M. Painho, Y. M. Santos, & H. Pundt (Eds.), *Geospatial thinking* (pp. 221–236). Berlin, Heidelberg: Springer. Retrieved from http://dx.doi.org/10.1007/978-3-642-12326-9_12

McDonough, L., Choi, S., & Mandler, J. M. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive Psychology*, *46*(3), 229–259.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Montello, D. R. (1993). Scale and multiple psychologies of space. In *Spatial information theory: A theoretical basis for GIS* (pp. 312–321). doi:10.1007/3-540-57207-4_21

Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's Downtown? Behavioral methods for determining referents of vague Spatial queries. *Spatial Cognition & Computation*, *3*(3), 104–185.

Piwek, P., Beun, R.-J., & Cremers, A. (2008). 'Proximal' and 'distal' in language and cognition: Evidence from deictic demonstratives in Dutch. *Journal of Pragmatics*, *40*(4), 694–718.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., & Yocum, Z. (2015). Semeval-2015 task 8: Spaceeval. In P. Nakov, T. Zesch, D. Cer, & D. Jurgens (Eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)* (pp. 884–894). Denver, CO: Association for Computational Linguistics.

Robinson, V. B. (2000). Individual and multipersonal fuzzy spatial relations acquired using human–machine interaction. *Fuzzy Sets and Systems*, *113*(1), 133–145.

Santos, J., Anastácio, I., & Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, *80*(3), 375–392.

Schockaert, S., de Cock, M., & Kerre, E. E. (2008). Location approximation for local search services using natural language hints. *International Journal of Geographical Information Science*, *22*(3), 315–336.

Schockaert, S., De Cock, M., & Kerre, E. E. (2009). Spatial reasoning in a fuzzy region connection calculus. *Artificial Intelligence*, *173*(2), 258–298.

Schockaert, S., De Cock, M., & Kerre, E. E. (2011). *Reasoning about fuzzy temporal and spatial information from the web* (Vol. 3). Singapore: World Scientific.

Shariff, A. R. B., Egenhofer, M. J., & Mark, D. M. (1998). Natural-language spatial relations between linear and areal objects: The topology and metric of English-language terms. *International Journal of Geographical Information Science*, *12*(3), 215–245.

Skoumas, G., Pfoser, D., & Kyrillidis, A. (2013). On quantifying qualitative geospatial data: a probabilistic approach. In D. Pfoser & A. Voisard (Eds.), *Proceedings of the Second ACM Sigspatial International Workshop on Crowdsourced and Volunteered Geographic Information* (pp. 71–78). New York, NY: ACM.

Stevens, A., & Coupe, P. (1978). Distortions in judged spatial relations. *Cognitive Psychology*, *10*(4), 422–437.

Wallgrün, J. O., Klippel, A., & Baldwin, T. (2014). Building a corpus of spatial relational expressions extracted from web documents. In C. B. Jones & R. S. Purves (Eds.), *Proceedings of the 8th Workshop on Geographic Information Retrieval* (pp. 6:1–6:8). New York, NY: ACM. Retrieved from http://doi.acm.org/10.1145/2675354.2675702

Wang, K., Thrasher, C., Viegas, E., Li, X., & Hsu, B.-j. P. (2010). An overview of microsoft web n-gram corpus and applications. In B. Settles, K. Small, & K. Tomanek (Eds.), *Proceedings of the NAACL Hlt 2010 Demonstration Session* (pp. 45–48). Stroudsburg, PA: Association for Computational Linguistics.

Williamson, T. (1994). *Vagueness*. London, UK: Routledge.

Worboys, M. F. (2001). Nearness relations in environmental space. *International Journal of Geographical Information Science*, *15*(7), 633–651.

Yao, X., & Thill, J.-C. (2005). How far is too far?–A statistical approach to context-contingent proximity modeling. *Transactions in GIS*, *9*(2), 157–178.

Zipf, G. (1935). *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.